Nikolaev Alexander

5130203/20102

In binary classification, we deal with an input space (space of instances) X and and output space (label space) Y. We identify the label space with the set {−1, +1}. The task involves assigning each object from space of instances to one of these two classes. The issue of learning can be simplified to estimating a functional relationship represented as *f*: X → Y. This type of mapping *f* is referred to as a classifier. In order to do this, we get access to some training points (X1, Y1), ...,(Xn, Yn) ∈ X × Y, drawn from an unknown probability distribution P(X, Y), the goal is to find function *f* that generalizes well to unseen data, minimizing misclassification errors. For this purpose, loss functions are used, for example the simplest loss function in classification is the 0-1-loss: $\ell(X, Y, f(X)) = \begin{cases} 1 & \text{if } f(X) \neq Y \\ 0 & \text{otherwise.} \end{cases}$

The risk of a function is the average loss over data points generated according to the underlying distribution P: $R(f) := E(\ell(X, Y, f(X))).$ In other words, the risk of a classifier f is the expected loss of the function f across all points X ∈ X. This risk measures the number of elements in the instance space X that are misclassified by the function *f*.

Of course, a function f is a better classifier than another function g if its risk is smaller, that is if R(f) < R(g). To find a good classifier f we need to find one for which R(f) is as small as possible. The best classifier is the one with the smallest risk value R(f). We can formally write down what the optimal classifier should be: $f_{Bayes}(x) := \begin{cases} 1 & \text{if } P(Y = 1 \mid X = x) \geq 0.5 \\ -1 & \text{otherwise.} \end{cases}$ This is the so-called “Bayes classifier”.

Statistical learning theory (SLT) provides a mathematical framework that is fundamental to understanding and solving binary classification problems in machine learning. In binary classification, SLT defines a set of functions that map input features to output classes (e.g., +1 or -1). This helps select the performance of a hypothesis from this space based on the trainee data. SLT introduces the concept of a loss function to quantify the failure of function *f*. The goal is to minimize the expected loss over the data distribution, resulting in better generalization ability.

The principle of empirical risk minimization suggests minimizing the empirical risk (the average loss over the trainee data) as an approach to finding good hypotheses. This balances good fit on the training data and the ability to deal with unseen data. SLT also provides theoretical bounds so that a well-learned hypothesis is known to work on unseen data.